Psychological Medicine, 1998, **28**, 915–921. Printed in the United Kingdom
© 1998 Cambridge University Press

Why GHQ threshold varies from one place to another

D. P. GOLDBERG,[1] T. OLDEHINKEL AND J. ORMEL

From the Institute of Psychiatry, London; and Department of Psychiatry, University of Groningen, The Netherlands

ABSTRACT

Background. No convincing explanation has been forthcoming for the variation in best threshold to adopt for the GHQ in different settings.

Methods. Data dealing with the GHQ and the CIDI in 15 cities from a recent WHO study was subjected to further analysis.

Results. The mean number of CIDI symptoms for those with single diagnoses, or those with multiple diagnoses, does not vary between cities. However, the best threshold is found to be related to the prevalence both of single and of multiple diagnoses in a centre. Variations in the diagnoses to be included in the 'gold standard' did not account for the variation observed. There was a strong relationship between area under the ROC curve (as a measure of the discriminatory power of the GHQ) and the best threshold, with higher thresholds being associated with superior performance of the GHQ. The items on the GHQ-12 that provided most discrimination between cases and non-cases varied from one centre to another.

Conclusions. The GHQ threshold is partly determined by the prevalence of multiple diagnoses, with higher thresholds being associated by higher rates of both single and multiple diagnosis. The mean GHQ score for the whole population of respondents provides a rough guide to the best threshold. In those centres where the discriminatory power of the GHQ is lowest, it is necessary to use a low threshold as a way of ensuring that sensitivity is protected, but the positive predictive value of the GHQ is then lower. Some of the variation between centres is due to variation in the discriminatory power of different items.

INTRODUCTION

In the original *Manual of the General Health Questionnaire* (Goldberg, 1978), variations in the best threshold to adopt were dealt with by saying 'there can be great difficulty about the most appropriate category for a borderline patient; so that a psychiatrist with a conservative concept of a psychiatric illness might use a higher threshold than a colleague with less clearly defined ideas. Neither is the psychiatrist the only variable: in other cultural settings, illness may be signalled by differing critical levels of symptom formation…' By the time the *User's Guide to the GHQ* arrived (Goldberg & Williams, 1988) the confusion was becoming more apparent, with the range of best thresholds for the GHQ-30 varying between 3/4 and 12/13; only the GHQ-60 had a relatively narrow range of best cutting thresholds. Apart from the presence of severe physical illness consistently raising the best threshold to be adopted, little sense could be extracted from the various thresholds proposed, and investigators were given advice on how best to determine the threshold in their particular cultural setting.

Van Hemert *et al.* (1995) took the argument further by demonstrating that free standing and disembedded versions of the GHQ had similar discriminatory ability (measured by area under

* Address for correspondence: Professor D. P. Goldberg, Institute of Psychiatry, De Crespigny Park, London SE5 8AF.

https://doi.org/10.1017/S0033291798006874 Published online by Cambridge University Press

the ROC curve), but differed in the threshold score where optimal discriminatory ability was obtained – with free standing versions of the GHQ requiring higher threshold scores than disembedded versions. A 'disembedded' version of the GHQ means that a shorter version of the test is extracted from a longer version by considering only the items that appear in the shorter test, 'free standing' means that the version of the GHQ used contains only that number of items. As one would have expected, they showed that more 'severe' concepts of caseness required higher threshold scores than less severe ones. This undoubtedly took the argument further, but did not solve the problem of variations in threshold when the same version of the GHQ is used against the same criterion of caseness.

An opportunity to study this phenomenon occurred with the WHO's study of *Mental Illness in General Health Care* (Ustun & Sartorius, 1995), where the GHQ-34, consisting of the items required for both the GHQ-12 and the GHQ-28, was used in 15 different cities, and validated against the Primary Care Version of the Composite International Diagnostic Interview (WHO, 1992). This does away with the 'differing concepts of psychiatric illness' by imposing a single (somewhat Western-orientated) criterion in 15 very different cities, leaving only variance due to the local population, and differing cultural expressions of psychological distress, to be accounted for. In an earlier paper we reported that variation in the best threshold to be adopted varied from a low of 1/2 to 6/7 for the GHQ-12, and from 3/4 to 7/8 for the GHQ-28 (Goldberg, *et al.* 1997). No effects were found for language (original language, English *versus* other languages), gender or educational level on the various validity coefficients considered. A review of 17 other validity studies using the GHQ-12 revealed an equally wide range of best estimates of the threshold score ranging from 0/1 to 5/6. Thus, the considerable variation in best thresholds is unaccounted for, and the purpose of the present paper is to study this variation in greater detail.

Five possibilities seemed worth exploring further: (1) that cities with higher thresholds might have disorders that were more severe, and this would manifest itself by cases of single disorders, and cases of multiple disorders, having higher symptom counts on the CIDI; (2) there might be a relationship between threshold and the prevalence of either single mental disorders, or multiple disorders ('co-morbidity') diagnoses by the CIDI; (3), that differing thresholds might be produced by different combinations of particular disorders in each city; (4), there might be a relationship between threshold and the overall discriminatory ability of the questionnaire; and (5), there might be differences in the complaint behaviour of people in different cities.

METHOD

The study involved 15 centres round the world, in which a total of 11 languages were spoken (Ustun & Sartorius 1995). Both the GHQ-12 and the primary care version of the Composite International Diagnostic Interview (CIDI-PC) were translated and back-translated in each of these languages. Consecutive patients attending clinics at participating centres were approached providing that they were older than 17, were not too ill to participate, were able to communicate and had a fixed address. The latter requirement was because the study used a longitudinal design.

A pilot study in each centre indicated that the GHQ score distributions were very different across centres. In each centre the scores were divided into three strata, so that the first contained approximately 60% of the patients, the second 20%, and the third the top 20%. To achieve these proportions, the following scores defined medium and high scorers in each centre: Ankara (2, 4), Athens (3, 5), Bangalore (3, 7), Berlin (2, 5), Groningen (2, 5), Ibadan (2, 5), Mainz (2, 5), Manchester (2, 4), Nagasaki (2, 4), Paris (4, 7), Rio de Janeiro (3, 5), Santiago (5, 7), Seattle (3, 5), Shanghai (0, 1 in one centre, 2, 4 in the other) and Verona (4, 6). The adoption of these varying sampling fractions meant that it was possible to complete the study in each centre by screening approximately similar numbers of patients. The GHQs were printed with a computer generated code indicating the score that patients should exceed if they were to be selected for interview: in this way a complex stratified sampling method was used smoothly across the 15 centres.

The CIDI-PC can generate diagnoses using either the International Classification of Disease, 10th Edition (ICD-10) system, or the Diagnostic

https://doi.org/10.1017/S0033291798006874 Published online by Cambridge University Press

and Statistical Manual of the American Psychiatric Association, 4th Edition (DSM-IV) system. For the present purposes, 'lifetime' diagnoses were ignored, and only current mental status was considered. The following diagnoses were included: current depression, agoraphobia, panic disorder, generalized anxiety disorder, neurasthenia (chronic fatigue) and mixed anxiety/depression. We examined the effects of including current anxiety symptoms (not requiring a 6 month duration) and alcohol dependence, but harmful use of alcohol was not included in the present analysis.

Sample size was determined in order to provide adequate statistical power for comparisons both between and within centres. It was projected that 1500 patients needed to be screened in each centre in order to detect 60 current cases of depression, and to have adequate numbers to allow a centre to compare the course of depression relative to other disorders that were common at that centre. If selected for interview, patients were usually seen within a day or so of the GHQ-12 being completed, although at Verona the time was within 2 weeks. At the time of the interview each subject completed a 34-item version of the GHQ, containing both the GHQ-12 and the GHQ-28. The present study has used results from this GHQ to compute validity measures. Thus, the GHQ-12 was administered on two occasions – the first time as the effective screening test to select subjects, and the second time on the day of the second-stage interview.

The research worker administering the CIDI-PC was blind to the results of this questionnaire. Selection of centres was dependent upon the existence of experienced investigators, and ability to raise funds for the study in the developed countries. The detailed methodology is described elsewhere (Ustun & Sartorius 1995).

RESULTS

Is threshold related to severity of illness?

The mean scores of non-cases, single diagnosis cases, and multiple diagnosis cases on the CIDI were not different in higher threshold centres compared with low thresholds centres (see Table 1). This shows that severity of illness as assessed by the CIDI is not affected by the threshold score on the GHQ-12. Centres with high thresholds have similar numbers of symptoms on the CIDI to those with low thresholds, and this applies both to single and to multiple disorders.

Table 1. *Mean scores on the CIDI sections of depression, anxiety and neurasthenia sections for non-cases, single diagnoses and multiple diagnoses*

Centre	Non-cases	Single disorders	Multiple disorders
Low threshold, 1/2			
Ankara	3·0	15·0	21·0
Ibadan	1·7	11·3	16·4
Nagasaki	1·0	9·9	17·4
Paris	2·6	12·4	20·0
Rio de Janeiro	3·9	15·4	24·1
Seattle	2·3	10·5	20·8
Shanghai	2·0	11·8	17·2
Verona	2·8	11·2	18·4
Mean for group	2·4	13·1	21·1
Mid-threshold, 2/3			
Athens	2·0	14·4	18·0
Berlin	3·2	10·7	15·9
Groningen	2·2	12·0	19·4
Mainz	3·0	11·3	18·8
Santiago de Chile	4·6	18·0	25·1
Mean for group	3·0	13·7	20·3
High threshold 3/4			
Manchester	1·6	12·5	17·5
High threshold 6/7			
Bangalore	1·5	12·6	19·6

Is threshold related to prevalence of disorder, or to multiple diagnoses?

The difference between mean number of disorders for low threshold (1/2) and higher thresholds, is significant ($t = -4{\cdot}98$, $P < 0{\cdot}001$, for unweighted data); the difference between prevalence of ICD-10 diagnoses for low and high threshold is also significant ($\chi^2 = 151{\cdot}75$, $P < 0{\cdot}001$, unweighted data). This relationship is fairly strong, in that the best threshold can be predicted in all centres except Berlin and Bangalore from a knowledge of the mean number of diagnoses or the proportion with multiple diagnoses. All low threshold centres have a mean number of diagnoses below 1·41 and a proportion with multiple diagnoses below 36%; while with the two exceptions all high threshold centres have a mean number of diagnoses higher than 1·46 and a proportion with multiple diagnoses above 38%. The relationship between the best threshold and the mean GHQ score for each population is quite

https://doi.org/10.1017/S0033291798006874 Published online by Cambridge University Press

Table 2. *Relationship between best threshold (on GHQ-12) and the mean GHQ score for the population; the prevalence of any of the disorders diagnosable by ICD-10 and the average numbers of different diagnoses per case, and proportion with multiple diagnoses*

City	Mean GHQ score	Prevalence of ICD diagnoses (%)	Mean number of disorders	Proportion with multiple diagnoses
Low threshold, 1/2				
Ankara	1·35	17·0	1·14	0·12
Ibadan	1·09	10·0	1·15	0·13
Nagasaki	1·12	9·9	1·25	0·17
Paris	2·14	28·3	1·41	0·32
Rio de Janeiro	2·32	33·4	1·41	0·36
Seattle	1·67	11·4	1·32	0·26
Shanghai	1·19	6·4	1·37	0·33
Verona	1·82	9·9	1·41	0·34
All the above	1·58	15·8	1·33	0·28
Mid-threshold, 2/3				
Athens	1·89	20·7	1·46	0·39
Berlin	2·56	19·8	1·38	0·35
Groningen	2·21	23·0	1·69	0·51
Mainz	2·11	22·0	1·48	0·38
Santiago	3·66	44·7	1·51	0·41
All the above	2·49	25·5	1·51	0·41
High threshold 3/4				
Manchester	2·78	27·3	1·66	0·48
High threshold 6/7				
Bangalore	3·03	15·4	1·39	0·34
All 2/3 or above	2·63	24·1	1·52	0·42

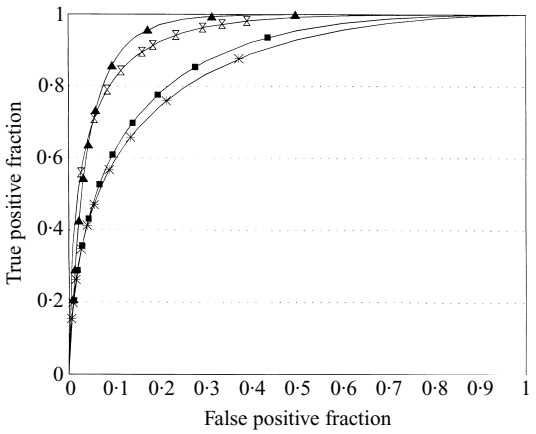

FIG. 1. Receiver operating curves (ROC) grouped by the best threshold found for the GHQ in each centre (weighted data). The size of the area under the curve (AUC) is a measure of the discriminatory power of the GHQ. (⚹, threshold 1/2; ■, threshold 2/3; ▲, threshold 3/4; ⧖, threshold 6/7.)

Table 3. *Relationship between best threshold on the GHQ and the overall discriminatory ability of the questionnaire, measured by the area under the ROC curve; the difference between* 2/3 *and* 3/4 *is significant;* 1/2 versus 2/3, *and* 3/4 versus 6/7, *are not*

Best threshold	Area under the ROC curve
1/2	0·85
2/3	0·87
3/4	0·95
6/7	0·94

strong, but still imperfect. This cannot be the whole story, as the differences between medium and high thresholds are not accounted for.

Could differences in the 'gold standard' account for different thresholds?

If this were so, we should expect varying the CIDI 'diagnoses' that are included in the notion of a 'case' would have an effect upon the threshold adopted. When best thresholds were calculated for each diagnosis separately, it was found that the threshold 2/3 was best for all of them, although for depression a threshold of 3/4 was equally good. With the exception of dysthymia in Bangalore, altering the diagnoses that were thought of as 'gold standard' – i.e. including or omitting current anxiety (rather than generalized anxiety) or dysthymic disorder, did not appear to affect the best threshold for the GHQ. In Bangalore, there was a high proportion of cases of dysthymic disorder, and these were classified as 'non-cases' in our original validity runs. Including dysthymic disorder as a diagnosis lowered the best threshold in Bangalore to 4/5, but had little other effect on the data.

Could best threshold be related to the overall discriminatory ability of the GHQ-12?

The relationship between best threshold on the GHQ-12 and overall discriminatory ability was computed by calculating the grouped area under the ROC curve for each of the four groups of centres (see Fig. 1). There was a steadily increasing area as threshold increased, and the differences between 2/3 and 3/4 were significant beyond the 0·01 level. It would appear that there is indeed a relationship between the best threshold and the overall performance of the GHQ-12 as a screening test, and in order to

https://doi.org/10.1017/S0033291798006874 Published online by Cambridge University Press

Table 4. *Table showing the PPVs and NPVs for each centre, both in the raw data, and in data corrected to the group prevalence of* 19·5%

City	Prevalence (%)	Observed values		Standard prevalence 19·5%	
		PPV%	NPV%	PPV%	NPV%
Threshold 1/2					
Ankara	17·0	45·0	93·2	49·1	92·0
Ibadan	10·0	29·6	97·0	47·8	93·7
Nagasaki	9·9	37·3	97·0	56·7	93·7
Paris	28·3	60·1	90·3	48·0	93·8
Rio de Janeiro	33·4	60·8	83·8	42·8	91·5
Seattle	11·4	31·0	97·1	45·8	94·6
Shanghai	6·4	22·7	98·8	50·9	50·9
Verona	9·9	19·2	96·0	34·5	91·5
Threshold 2/3					
Athens	20·7	57·9	94·4	56·1	94·7
Berlin	19·8	41·8	91·7	41·3	91·9
Groningen	23·0	63·8	93·6	58·9	94·8
Mainz	22·0	52·4	91·6	48·6	92·7
Santiago de Chile	44·7	79·4	86·9	53·5	95·7
Threshold 3/4					
Manchester	27·3	74·8	93·9	65·7	96·0
Threshold 6/7					
Bangalore	15·4	58·7	97·3	65·4	96·5

investigate this further the positive predictive value (PPV) and negative predictive value (NPV) of the GHQ were calculated for each centre separately. Both of these indices are highly dependent upon prevalence, so in order to disentangle the effects of prevalence and the effects of differing thresholds, we re-calculate the value for each city at the prevalence of disorder in the whole data-set, which was 19·5%.

It can be seen from Table 4 that the considerable differences between centres become much less severe in the right hand columns, with prevalence controlled. However, as one would expect, those centres using a low threshold have to tolerate very much poorer PPVs than those with higher thresholds.

Is the GHQ-12 responded to in the same way in different places?

The items in the GHQ that account for its discriminatory ability were considered in two ways, by calculating the three most specific items (those items with the lowest scores in non-cases), the three most sensitive items (with the highest scores among cases), as well as overall discriminatory ability as judged by the results of a multiple logistic regression analysis.

It can be seen from Table 5 that very few items behave in the same way in each city: only items 10, 11 and 12 are very similar, with 1 and 2 following behind. The other 7 items are really highly variable in the way that they are responded to in each city, although no item can

Table 5. *Items from GHQ-12 in 15 centres, showing how often each item was among the three most sensitive items for each city; the three most specific; or was chosen in the multiple regression analysis as a discriminant item for each city**

GHQ-12 item	3 most specific	3 most sensitive	Multiple regression†
1 Lost sleep	0	8	4
2 Under strain	1	10	5
3 Lost concentration	1	5	3
4 Play a useful part	9	0	1
5 Face problems	3	1	1
6 Make decisions	7	0	4
7 Overcome difficulties	4	1	2
8 Felt happy	1	6	1
9 Enjoy activities	1	5	5
10 Depressed	1	12	9
11 Losing confidence	11	1	4
12 Felt worthless	12	0	6

* Thus, the '11' for '3 most specific' for the item GHQ-11 means that this item was among the 3 most specific items in 11 of the 15 cities.

† The 3rd column considers the 12 items as a group in each city, and calculates the independent contribution of each item to the overall discrimination between cases and normals: the '4' for GHQ-11 indicates that this item was among the three best discriminant items in 4 of the 15 cities.

https://doi.org/10.1017/S0033291798006874 Published online by Cambridge University Press

be discarded as there is somewhere in the world where it appears to be a useful item. Thus, genuine differences in the way in which distress is experienced from one place to another account for differences in the mean for the whole population, and the number of different diagnoses experienced by the subjects: each of these in part determines the threshold which will be best for that population.

DISCUSSION

The study is unusual in having almost identical procedures in 15 very different cities. The exception was the GHQ being read out to respondents in Bangalore. Thus, variance due to the research interview is effectively removed.

Anomalous results

The high threshold in Bangalore appears to be due to three factors: it is partly because the single commonest disorder in Bangalore was dysthymic disorder (weighted prevalence 9·9%; all disorders weighted prevalence 23·9%). We have seen that the high score was because dysthymic disorder had been declared to be a non-case, and we would expect this to raise the best threshold considerably: reassigning dysthymic disorder to the 'case' category lowered the best threshold to 4/5. The second factor is that the questionnaire was read out to all respondents in Bangalore, as so many were illiterate: this may well influence the responses of the subjects. However, a third factor is that the GHQ seems to work well as a discriminator in this centre – and this allows a higher threshold to be preferred for the reason already given. The anomalous position of Berlin in Table 2 can be partly understood by the finding that it does almost as well as a 1/2 centre (2/3 sensitivity 72·6%, specificity 75%; 1/2 sensitivity 82·2%, specificity 70·0%).

The original hypotheses

(i) The demonstration that the mean number of CIDI symptoms for either single or multiple diagnoses is not affected by best threshold is an important one, since it indicates that high thresholds are not because cases are more severe in some places than others.

(ii) The relationship between multiple diagnoses and best threshold indicates that high thresholds are associated with what is sometimes referred to as 'co-morbidity' and that this varies from one place to another. The relationship between threshold and prevalence might also be caused by the same phenomenon, since as prevalence increases the number of cases of multiple diagnosis also increases. The higher GHQ means found with greater thresholds reflects the effects of higher rates for both single and multiple diagnoses. It seems unlikely that high prevalence of both single and multiple disorders in Bangalore and Manchester accounts for the better discrimination obtained in those centres, as Rio de Janeiro and Santiago de Chile had higher prevalences, but lower thresholds.

(iii) The finding that similar best thresholds were required for individual diagnoses considered on their own is strong evidence against the idea that different combinations of disorders accounted for the varying thresholds.

(iv) There does indeed appear to be a relationship between best threshold and the overall discriminatory ability of the GHQ, with those centres with a higher threshold enjoying a higher discriminatory ability (Fig. 1). There are two reasons why this might be so. The first is that centres with high threshold tend to have high prevalence (Table 2), and with higher prevalence the PPV is bound to be better: but if this were the only explanation, the NPV would be correspondingly worse – and we have seen from Table 4 that this is not so. It would appear that in those centres where the GHQ-12 discriminates best between cases and normals it is possible to use a higher threshold without sacrificing sensitivity. The higher threshold naturally achieves a superior specificity. In the study of van Hemert *et al.* (1995), there was no relationship between the size of the area under the ROC curve and ascending thresholds: but the subjects were all drawn from the same culture, whereas the present study examines differing thresholds using the same questionnaire in different cultures. Thus, there is no conflict between the Dutch results and our own, except that van Hemert *et al.* report a best threshold for the free standing GHQ-12 of 5/6 against the PSE > 5 in Leiden, whereas we found it to be only 2/3 against the CIDI in Groningen. (Many investigators have used a lower criterion of caseness than > 5, and one would expect that the more usual 4/5 might produce a lower threshold.)

https://doi.org/10.1017/S0033291798006874 Published online by Cambridge University Press

(v) The remaining differences between centres are likely to be accounted for by variations in the way that individual items are responded to, or the 'complaint behaviour' of the respondents (Table 5). The discriminatory power of individual test items varies from place to place, and thus the thresholds will be affected to some extent. Language is unlikely to be the key to better discrimination – as Seattle had a lower threshold than Manchester, and the GHQ was administered in Bangalore in Kannada, not English.

Implications for future epidemiological research

For those wishing to achieve an optimal trade-off between sensitivity and specificity, it remains the case that carrying out one's own validity study, as recommended in the User's Guide, is the safest option. However, the mean score found in a pilot study will provide a rough guide to the best threshold (see Table 2), and will not be far from that found with a more expensive study. In primary care settings, if the mean is below 1·85 then a threshold of 1/2, from 1·85 to 2·7 a threshold of 2/3, and above 2·7 a threshold of 3/4 seems to work best for the GHQ-12. Knowledge of the proportion of cases with multiple diagnoses will add very little to this rule of thumb. In places where the discriminatory power of the GHQ is lowest, it is necessary to use a low threshold as a way of ensuring that sensitivity is protected: the cost of using a low threshold is that the positive predictive value of the GHQ is lower in such centres.

REFERENCES

Goldberg, D. P. (1978). *Manual of the General Health Questionnaire*. NFER: Windsor.

Goldberg, D. P. & Williams, P. (1988). *A User's Guide to the General Health Questionnaire*. NFER: Windsor.

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O. & Rutter, C. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine* **27**, 191–198.

Ustun, B. & Sartorius, N. (1995). *Mental Illness in General Health Care. An International Study*. John Wiley: Chichester.

van Hemert, A. M., Den Heijer, M., Vorstenbosch, M. & Bolk, J. H. (1995). Detecting psychiatric disorders in medical practice using the General Health Questionnaire: why do cut-off scores vary? *Psychological Medicine* **25**, 165–170.

https://doi.org/10.1017/S0033291798006874 Published online by Cambridge University Press